INTUITIVE THEORIES OF INFORMATION: BELIEFS ABOUT THE VALUE OF REDUNDANCY

by

J. B. SOLL*

98/01/TM

* Assistant Professor of Decision Sciences at INSEAD, Boulevard de Constance, 77305 Fontainebleau Cedex, France.

TECHNOLOGY MANAGEMENT

SOLL Jack B.:
Intuitive Theories of Information: Beliefs About the Value of Redundancy.
INSEAD N° 98/01/TM.

In many situations, quantity estimates from multiple experts or diagnostic instruments must be collected and combined. Normatively, one should value information sources that are both accurate and nonredundant (i.e., one should minimize correlation in forecast errors). Past research has produced conflicting results with respect to preferences for redundant information. This paper shows that this preference depends on the interaction between one's intuitive theory of information and one's beliefs about the situation-specific error generating process. Manipulations of the perceived source of error (e.g., measurement error vs. systematic bias) lead people to prefer more or less redundancy in predictable ways. Additionally, science training is associated with having an intuitive theory that matches the normative model, whereas statistical training is not.

Running Head: BELIEFS ABOUT REDUNDANCY

Intuitive Theories of Information:

Beliefs about the Value of Redundancy

Jack B. Soll

INSEAD
Fontainebleau, France

Beliefs About Redundancy

Jack B. Soll
INSEAD
Bd. de Constance
77305 Fontainebleau Cedex
France

phone: (33) 1 60 72 44 49
email: jack.soll@insead.fr

Abstract

In many situations, quantity estimates from multiple experts or diagnostic instruments must be collected and combined. Normatively, one should value information sources that are both accurate and nonredundant (i.e., one should minimize correlation in forecast errors). Past research has produced conflicting results with respect to preferences for redundant information. This paper shows that this preference depends on the interaction between one's intuitive theory of information and one's beliefs about the situation-specific error generating process. Manipulations of the perceived source of error (e.g., measurement error vs. systematic bias) lead people to prefer more or less redundancy in predictable ways. Additionally, science training is associated with having an intuitive theory that matches the normative model, whereas statistical training is not.

In many situations, multiple estimates of an uncertain quantity must be collected and combined into a single aggregate estimate. Consider, for example, an oil company that is deciding where to drill for oil. The company will base this decision in part on the opinions of professional geologists. The very first question that the company faces, then, is how many geologists to consult, and which specific ones. Only after the geologists have done their work can the company advance to the complex task of combining multiple opinions. Of course, there are many variations on this theme. The geologists could arrive at a group consensus, for instance, and hence solve the aggregation problem on their own. Whatever the specific process, we can speak of a general problem in which multiple estimates or opinions must be collected and then somehow combined into a final judgment upon which to base decisions (Wallsten, Budescu, & Erev, 1997).

Because of its importance and widespread application, this general problem has drawn attention from both theoreticians and applied researchers. Curiously, most of the effort has gone toward understanding the aggregation phase of the problem. An important empirical result is that "merely" averaging multiple estimates is a remarkably effective way to reduce forecast error (Clemen, 1989; Ferrell, 1985; Zajonc, 1962). The optimal number of estimates to include in a composite is generally between six and twenty (Ashton, 1986; Hogarth, 1978), and usually most of the benefit accrues with just the first two or three (Libby & Blashfield, 1978; Makridakis & Winkler, 1983). Some successful applications of averaging include clinical judgment (Goldberg, 1965; Overholser, 1994), macroeconomics (Clemen & Winkler, 1986), business (Larréché & Moinpour, 1983), and meteorology (Sanders, 1963; Staël Von Holstein, 1971). Meanwhile, psychologists have examined how people intuitively combine

opinions, either as individuals (Birnbaum & Stegner, 1979), or in groups (Einhorn, Hogarth, & Klempner, 1977; Hastie, 1986). The final judgment is typically modeled as a weighted-average of the inputs (but see Sniezek & Henry, 1990), although recent research has emphasized the psychological process leading up to the weights rather than the weights themselves (Heath & Gonzalez, 1995; Sniezek & Buckley, 1995).

In contrast to aggregation, estimate collection has received little attention, especially from psychologists. Normatively, both accuracy and redundancy are important considerations in choosing and aggregating opinions. The accuracy of an average depends on the accuracies of the individual inputs, and also on the extent to which the inputs are independent. To see the point about independence, consider three equally competent analysts who forecast prices. Analysts A and B work together, observe similar events, and consult one another frequently, while Analyst C works alone. The opinions of Analysts A and B are probably redundant, in the sense that their forecast errors are likely to be highly correlated. Having purchased A's opinion, one might do better to approach C rather than B. The same principle applies in scientific measurement. Given two equally valid machines that measure radiation, it makes more sense to take one reading from each machine rather than two from the same machine. Although the redundant judgment or measurement has value, the nonredundant source dominates due to the lower expected correlation in errors. In general, the use of similar machines or methods, common experience, and frequent exchange of opinion can all lead to shared biasing characteristics, and hence to redundant estimates (Hogarth, 1989; Stasson & Hawkes, 1995).

The present paper explores how people collect opinions and estimates, and thus begins to fill a gap in the psychological literature on opinion aggregation. I deal here exclusively with the redundancy issue, leaving tradeoffs between accuracy and redundancy for future research. The reason for this scope is that the few past studies of redundancy disagree about whether people put positive or negative value on correlation. For example, Goethals & Nelson (1973) asked students to predict the academic performance of potential new students on the basis of videotaped interviews. The students were more confident in their final predictions when they learned that a peer with a dissimilar judgment style agreed with their initial prediction, as opposed to a peer with a similar style. Presumably, people with similar judgment styles will have correlated forecast errors. An aggregate based on multiple styles can be expected to be more accurate, and therefore warrants greater confidence (Mettee & Smith, 1977). The result implies that people correctly put negative weight on redundancy. In contrast, Kahneman & Tversky (1973) reported that people are more confident predicting final grade point averages from highly correlated cues, such as two science grades, as opposed to less correlated cues, such as one science and one English grade (see also Slovic, 1966). They concluded that people erroneously assume that consistency implies validity, when in fact consistency is often a product of correlated inputs. This implies that people put positive weight on redundancy. Several studies of opinion weighting (Gonzalez, 1994; Russ, Gold, & Stone, 1979, 1980) support Goethals and Nelson's conclusion, while another suggests that people are indifferent about redundancy (Maines, 1996).

Past work on redundancy suggests two things. First, there is a potentially interesting story to be told about how people think about redundancy. Second, any model of how people make tradeoffs between accuracy and redundancy is going to be very complicated, since the directional valuation of redundancy shifts across studies. This paper is a first attempt at describing how people reason about redundancy when choosing sources of information. I will first describe several exploratory studies that replicate the contradictory preferences of past research. Next, Experiment 1 shows that whether people value redundancy positively or negatively depends upon perceptions of the source of estimation errors in the problem at hand. This helps explain the contradictory results of past studies. Finally, Experiment 2 describes people's intuitive theories of what happens to estimation errors when multiple estimates are combined. Preferences for redundancy in a particular situation can be traced to these intuitive theories.

Exploratory Studies

On the whole, the existing evidence gives no clear a priori reason to expect people to prefer redundant or nonredundant information when given a direct choice. Given the paucity of data on this issue, two preliminary studies were carried out to see if there is a general preference (Soll, 1997). In the first, participants used case specific information to predict future bankruptcy ratings of new loan applicants, and also to assess the chances that sexual harassment cases would be decided in favor of the defendant. Participants reviewed ten past cases with outcomes, predicted outcomes for new cases, and finally revised predictions with the aid of a peer advisor's predictions for the same cases. Participants chose between a peer who reviewed the same ten past cases as they did in the first stage of the study, or one who

reviewed a different set. Payment was based on the accuracy of the final judgments. As participants had no reason to believe that either advisor had greater ability, normatively they should have chosen the nonredundant one. All 40 participants went through this routine twice, once for the bankruptcy task and once for the legal task. They could thus choose the redundant advisor 0, 1, or 2 times. The corresponding frequencies were 13, 15, and 12, respectively.

The above result could just reflect random behavior; people may be insensitive to redundancy considerations. Alternatively, people may differ in their general preferences for redundant or nonredundant information sources, with roughly half coming down on each side. A final possibility is that people see advantages to both types of information sources, and make tradeoffs in choosing between the two. A second preliminary study attempted to discriminate among these explanations by asking people to explain their preference. Survey respondents were given a hypothetical scenario in which a hospital administrator collects and averages two expert judgments of the percentage of a patient's liver affected by cancer. Doctors at the hospital use one of two equally diagnostic methods to diagnose liver cancer. Respondents were asked whether they favored consulting two doctors who use the same method or two who use different methods, and to describe their reasoning in short essays. Seventeen of the thirty-three respondents chose two doctors who use the same method. While there is no clear preference, the written explanations are illuminating. The following comments are typical of those preferring the same method.

Since both tests are equally valid, using the same test twice will be more likely to catch the errors within that test, whereas using both tests won't reflect anything except the differences between the tests.

Since A and B are regarded equally, they are presumably equally valid. However, since there is no information on how A and B differ in their errors (tendency to estimate high, low, etc.) it is more dangerous to try to mix them than to take one or the other. Also, the 2 doctors would be experts but there is a better chance they will catch each other's mistakes if they agree on a procedure.

If looking for a reliable number you'd want to use the same method twice so you eliminate an uncontrollable variable.

The above respondents recognize that a given method is prone to error on any given application, and are tempted to use the same method twice to reduce this within-method error. In contrast, those choosing both methods explained that this protects against the idiosyncratic bias of a given method.

In case some strange effects caused one method to give slightly different results, using one of each would help.

> Two methods not subject to the same possible flaws are more likely to average out to a more precise estimate.
>
> The two tests most likely complement each other. Test A catches things Test B misses and vice versa.
>
> More methods are better, and this should eliminate (hopefully) systematic errors in the tests.

Regardless of preference, the above quotes reveal a high degree of statistical sophistication. People recognize two kinds of error, nonsystematic within-method errors and systematic between-method errors. They also see the benefit of reducing each type of error. Normatively, using nonredundant sources reduces both types of error simultaneously. In contrast, many people seem to believe that using nonredundant sources reduces only systematic error (this is false), and using redundant sources reduces only nonsystematic error (this is true). In other words, people apparently see conflict, or a tradeoff, where none exists. I call this general intuitive model of error reduction the Error Tradeoff Model (ETM). A belief in ETM implies that people will first partition error into nonsystematic and systematic components, and then determine which is more important for the problem at hand. A natural implication of ETM is that people will prefer redundant sources when they perceive that nonsystematic error contributes more to total error, and nonredundant sources when they

perceive that systematic error contributes more. Experiment 1 tests this prediction by manipulating perceptions of the two types of error.

Experiment 1

Anecdotally, many participants struggle with thought problems like the medical problem above, and mention that there are good reasons for either action. However, the explanations in the preliminary studies could reflect ex post rationalization, rather than the actual mental representations and mechanisms that underlie choice. With this in mind, the present experiment attempts to manipulate, between-subjects, the perceived relative sizes of systematic and nonsystematic errors. ETM individuals should tend to prefer nonredundant information sources when systematic errors appear large, and redundant sources when nonsystematic errors appear large.

Materials

Scenarios were constructed for two disparate domains of knowledge. In both, an information seeker has already consulted one source, and now must choose between two others to obtain a more precise estimate of the truth. Problem 1 involves seeking opinions from experts who observe perceptual stimuli. Problem 2 involves the reading of diagnostic tests that are subject to random fluctuation and systematic bias.

Problem 1:

> Imagine that you are a field commander in the midst of a difficult ground war. An opposing army rests in a valley 15 miles ahead, and you need an estimate of its size. You can send a scout to one of two equally good vantage points, Agnon Cliffs or Wilbur's Peak. Suppose you send

a scout to Agnon Cliffs under stormy (sunny) conditions. The scout's best guess is 9,000 troops, and based on this report you think that the true number is somewhere between 6,000 and 12,000. To improve your estimate you decide to send a second scout. The weather is now sunny at both locations. Assuming that the scout will return safely, where would you send him?

Agnon Cliffs Wilbur's Peak

Problem 2:

Imagine that two equally accurate home kits for measuring blood cholesterol have arrived on the market, each costing $20 and good for one use. Both brands of kits can make mistakes. If you use a given brand over and over again, you will typically notice somewhat different (very similar) readings. If you use each brand once, you might find a larger difference in the readings, because they use different chemical processes. You try Brand A on a family member, and the reading is 180. You decide to buy another kit, and to base your final estimate on the results of both tests. Given that you want your final estimate to be as close to your relative's true cholesterol level as possible, would you try Brand A again, or Brand B?

In Problem 1, the word stormy in the fourth sentence was changed to sunny for half the participants. In Problem 2, the words somewhat different were changed to very similar. Both of these manipulations are likely to affect how total error partitions into systematic and

nonsystematic components. Stormy weather tends to blur the visual field, making it more likely that a larger portion of the total uncertainty comes from random perceptual errors. In Problem 2, the relative size of the nonsystematic component is conveyed explicitly. Nonsystematic error is high when within-test readings are somewhat different, and low when they are very similar. One difference between the two problems is that Problem 1 holds constant the total amount of uncertainty across conditions, whereas Problem 2 may not. However, in both problems the manipulation should affect the perceived ratio of systematic to nonsystematic error. If people believe in ETM, they should be more likely to prefer nonredundant information sources the greater this perceived ratio.

Procedure and Results

Participants were 402 students at Northwestern University and the University of Chicago. Each participant answered one version of either Problem 1 or 2. When weather conditions in Problem 1 were stormy, 58% preferred Agnon Cliffs, compared to 42% when it was sunny ($N = 200$, Yates' $\chi^2 = 4.50$, $p < .05$). In Problem 2, 63% preferred Brand A when within-brand results were somewhat different, compared to 48% when they were very similar ($N = 202$, Yates' $\chi^2 = 4.51$, $p < .05$).

Discussion

The present results confirm that many people prefer to consult redundant sources. The results also support the ETM hypothesis. According to ETM, the nonsystematic error due to the storm at Agnon Cliffs can only be reduced by going back to Agnon Cliffs. Similarly, the trial-to-trial error of a noisy Brand A can only be reduced with repeated uses of Brand A. When nonsystematic or trial-to-trial error appears larger, people are more likely to use

redundant information sources. The manipulations in Problems 1 and 2 are very subtle, and therefore this study was expected to reveal only a portion of those participants who follow ETM. For example, some participants may have believed that the bulk of the error in Problem 1 is systematic, regardless of weather conditions. These individuals would anticipate slightly more nonsystematic error on a stormy day, but in both conditions would select Wilbur's Peak. Overall, in support of ETM the manipulation changed perceptions enough to affect preferences for approximately 8% of respondents in Problems 1 and 2. However, ETM predicts a preference to use both brands in Problem 2 when repeated uses give very similar results. In fact, only 52% preferred to use both. It appears that ETM explains the choices of some but not all participants. In the following sections, a framework is developed for describing ETM and alternative intuitive theories.

Normative Theory

I consider here only those cases in which an information seeker arrives at a final judgment by averaging (either weighted or simple) two quantity estimates. In some situations, the information seeker will have an initial opinion that bears on the final judgment. If so, then for present purposes this counts as one of the two estimates. The relationship between the two estimates and the true criterion can be expressed by

$$\underline{X}_1 = \underline{T} + \underline{e}_1 \qquad (1)$$

$$\underline{X}_2 = \underline{T} + \underline{e}_2 , \qquad (2)$$

where $\underline{T}$ represents the realized value of the criterion, $\underline{X}_1$ and $\underline{X}_2$ are the two estimates obtained by the information seeker, and $\underline{e}_1$, $\underline{e}_2$ are mean-zero error terms associated with these estimates. The information seeker averages $\underline{X}_1$ and $\underline{X}_2$ to obtain $\underline{X}_C$, the final judgment. This

formulation does not account explicitly for bias. Regarding bias, there are two possibilities. First, the information seeker may believe that an information source could be biased, but has no information as to the direction of the bias. In this case, the uncertainty in the bias can be incorporated directly into the mean-zero error term. Second, the information seeker may believe that a source is biased in a certain direction. Here, the information seeker can subtract out the expected bias of the source, and use this bias-adjusted estimate as the input in calculating $\underline{X}_C$. Either way, the relationship between the final judgment and the criterion is expressed as

$$\underline{X}_C = \underline{T} + \underline{e}_C , \qquad (3)$$

where $\underline{e}_C$ is mean-zero. I assume that the information seeker wishes to minimize σ_C^2, the variance of $\underline{e}_C$. Within the present framework, two factors are relevant to minimizing σ_C^2: (1) the accuracies of the two estimates that combine to produce $\underline{X}_C$; and (2) the extent to which the errors associated with these two estimates are correlated. The accuracies of the individual estimates are given by σ_1^2 and σ_2^2, which are the error variances of $\underline{e}_1$ and $\underline{e}_2$, respectively. Typically, the correlation between $\underline{e}_1$ and $\underline{e}_2$ will be higher to the extent that the two estimates come from similar information sources. Barring certain well-defined anomalous circumstances and ceteris paribus, one should choose information sources so as to minimize the correlation between the errors of the two estimates.[1]

One "right" way to think about redundancy is to recognize that positive correlation in errors is undesirable and that dissimilar information sources will tend to produce less correlated errors. Experiment 1, however, showed that everyday thinking does not readily map into acceptance or rejection of this normative rule, as at least some people partition error

into systematic and nonsystematic components. To show precisely how people's beliefs about error reduction may differ from the normative model, it is first necessary to show how statistical theory deals with multiple sources of error. I begin by assuming that the two error types are independent. This seems reasonable given that nonsystematic error is often measurement or perceptual error. Given independence, the two error types are separable in a decomposition of variance.

Let V_j and V_k be the "true readings" from information sources j and k, respectively, where j and k index the n information sources to which the decision maker has access ($j, k = 1, ..., n$). The term j indicates the information source associated with estimate X_1, and k the source associated with X_2. The two estimates are from the same information source if $j = k$ and from different sources if $j \neq k$. By definition, V_j and V_k are the estimates that sources j and k would produce on a repeated basis were nonsystematic error not a factor (for a similar approach, see Erev, Wallsten, & Budescu, 1994). Transient environmental factors and random judgmental error might both cause deviations from the true reading, and produce nonsystematic error. The error variances can be decomposed into systematic and nonsystematic components by adding and subtracting V_j.

$$\begin{aligned} \sigma_1^2 &= E(X_1 - T)^2 \\ &= E[(X_1 - V_j) + (V_j - T)]^2 \\ &= E(X_1 - V_j)^2 + E(V_j - T)^2 + 2E[(X_1 - V_j)(V_j - T)] \\ &= E(X_1 - V_j)^2 + E(V_j - T)^2 + 2E(X_1 - V_j)E(V_j - T) \end{aligned} \quad (4)$$

The last equality holds because the nonsystematic and systematic errors are independent. I assume that from an ordinary person's point of view, nonsystematic and systematic errors are

both mean-zero. As discussed earlier, this does not mean that people believe that the information sources are unbiased, but rather that the they have no idea about the direction of this bias. From a subjective perspective, bias is a mean-zero random variable. Equation 4 then reduces to

$$\sigma_1^2 = \underline{E}(\underline{X}_1 - \underline{V}_j)^2 + \underline{E}(\underline{V}_j - \underline{T})^2 , \qquad (5)$$

The decomposition of σ_2^2 is similar. To simplify the notation, I will use $\underline{S}$ and $\underline{N}$ to represent the systematic and nonsystematic portions of error variance, respectively. That is

$$\underline{S}_1 = \underline{E}(\underline{V}_j - \underline{T})^2$$

$$\underline{S}_2 = \underline{E}(\underline{V}_k - \underline{T})^2$$

$$\underline{N}_1 = \underline{E}(\underline{X}_1 - \underline{V}_j)^2$$

$$\underline{N}_2 = \underline{E}(\underline{X}_2 - \underline{V}_k)^2 .$$

The error variance of an estimate is simply the sum of the systematic and nonsystematic components. Consider what happens to total error variance when a final judgment is based on two estimates. Here, I consider the simple case in which $\underline{X}_1$ and $\underline{X}_2$ have equal error variances. Given that the two information sources are equally valid, the best strategy is simply to allow one's final combined estimate, $\underline{X}_C$, to equal the average of the two individual estimates $\underline{X}_1$ and $\underline{X}_2$. The error variance of $\underline{X}_C$ is as follows, where $\underline{\rho}_S$ is the correlation between the two systematic errors.[2]

$$\underline{E}(\underline{X}_C - \underline{T})^2$$

$$= \frac{\underline{N}_1+\underline{N}_2}{4} + \frac{\underline{S}_1+\underline{S}_2}{4} + \frac{1}{2}\sqrt{\underline{S}_1\underline{S}_2}\,\underline{\rho}_S$$

$$= \underline{N}_C + \underline{S}_C$$

The derivation of the first equality is straightforward (see Appendix A). $\underline{N}_C$ represents that portion of the final error variance due to the nonsystematic errors of the individual estimates, and $\underline{S}_C$ represents that portion due to systematic errors. The intuition is easiest to follow for the special case in which both information sources have equal amounts of systematic and nonsystematic error (assuming otherwise does not affect the normative prescription). In that case, let $\underline{N} = \underline{N}_1 = \underline{N}_2$ and $\underline{S} = \underline{S}_1 = \underline{S}_2$. Then

$$\underline{N}_C = \frac{\underline{N}}{2}$$

$$\underline{S}_C = \frac{\underline{S}}{2}(1+\rho_{\underline{S}})$$

These equations clarify why using nonredundant sources dominates. Nonsystematic error is reduced by half whichever information sources are consulted. In contrast, the degree to which systematic error is reduced is moderated by $\rho_{\underline{S}}$. If the same information source is consulted twice, $\rho_{\underline{S}} = 1$ and systematic error stays at $\underline{S}$. If different information sources are consulted, $\rho_{\underline{S}} < 1$, and systematic error is reduced. The appropriate beliefs about error reduction, then, can be stated succinctly as follows:

If information sources are redundant,

$$\underline{N}_C < \underline{N}$$

$$\underline{S}_C = \underline{S}$$

If information sources are nonredundant,

$$\underline{N}_C < \underline{N}$$

$$\underline{S}_C < \underline{S}$$

Both redundant and nonredundant sources reduce nonsystematic error. Nonredundant sources are more valuable, however, because they also reduce systematic error.

Modeling Beliefs

The above error-partitioning framework can be used to describe a variety of possible intuitive theories. In the current set of experiments, participants are given a choice between two readings from the same information source (redundant) and two readings from different sources (nonredundant). An intuitive theory is a set of beliefs about what happens to the two error types when a given information collection strategy is used. Since there are two strategies (redundant and nonredundant), an intuitive theory includes four specific beliefs. Table 1 shows the intuitive theories that correspond to the Normative Model and to ETM. The primary feature of ETM is that no information search strategy reduces both nonsystematic and systematic errors, and therefore in choosing a source one makes a tradeoff. Table 1 indicates that there are several variants of ETM, reflecting different beliefs about what happens to the error types that are not being reduced. Interestingly, if Relation 2 holds as an equality then ETM and the Normative model differ only in Relation 3. This one difference will generate a preference for redundancy whenever nonsystematic error is perceived as more important. Notice also that each of the four error relations can take on three values, implying a total of 81 possible intuitive theories. These can sometimes be grouped into larger, more meaningful categories, such as ETM. Additional intuitive theories will be discussed in the context of Experiment 2.

The exploratory studies and Experiment 1 show that people frequently prefer redundant information sources, and that intuitive theories play a role in this preference.

Experiment 1 further shows that ETM underlies choice for at least some people. Yet several questions remain. Just how common is ETM as an intuitive model? Do people vary in the intuitive models that they hold, or do they tend to cluster around just a few of them? Experiment 2 answers these questions by providing a fuller map of the beliefs and rules that guide information search behavior. The primary goal is to pare down the list of 81 possible intuitive theories to just several, and to categorize people accordingly.

Experiment 2

This experiment uses multiple scenarios that explicitly inform participants about the potential for systematic and nonsystematic errors. The advantage of this approach is that, unlike Experiment 1, perceptions of the relative contribution of each error type will not vary across participants. This enables an analysis wherein a participant's set of responses across scenarios can be used to infer a specific intuitive theory.

The stimuli describe technicians who use scales to weigh very small objects. Each scale has both bias (systematic error) and measurement error (nonsystematic). There are three scenarios overall, each of which describes the scales and procedures of a different laboratory. The amount of possible bias and measurement error vary across the scenarios. The design is completely within-subjects. In each scenario, participants learned about two technicians: one who averages two measurements from the same scale, and another who averages two measurements from different scales. Participants judged which technician, over the long run, would be more accurate.

The study measures beliefs in three ways. First, beliefs about the four error relations (see Table 1) are inferred indirectly from accuracy judgments. Each participant produced

three accuracy judgments altogether, one for each scenario, and in many cases this pattern is consistent with only a few sets of beliefs. Second, participants explained their judgments in writing. Third, participants were asked explicitly to indicate their beliefs about what happens to each type of error when estimates are combined. A correspondence between inferred and reported beliefs would indicate that the experiment successfully taps and describes the mental representations that guide search behavior. A lack of correspondence would possibly reflect unstable beliefs, or behavior guided largely by inaccessible rules and scripts (Nisbett & Wilson, 1977).

Method

Participants.

Fifty-seven University of Chicago students were paid $6 each; most took roughly 30 minutes to complete the task. Two students were later excluded from the analysis because their responses indicated that they did not fully understand the instructions.

Materials and Procedures.

All materials were included in a single booklet. The first page described the task and provided basic definitions. Participants were told that they would evaluate the procedures of technicians who weigh small objects. They were informed that measurement error reflects the fact that a given scale typically registers different readings each time it weighs the same object, and that bias reflects the fact that a given scale tends to over- or underestimate the weights of all objects. The term total deviation was introduced, with examples, as the absolute difference between the reading on the scale and the actual weight of the object. Participants rated the terms bias, measurement error, and total deviation on a 9-point scale,

with the endpoints labeled Extremely hard (Extremely Easy) to understand. The three scenarios were then presented sequentially, with the order completely counterbalanced across participants. Each scenario discussed technicians at the Calon, Kensington, or Alpine lab. In each lab, one technician always takes two measurements from the same scale for each object and records the average. The other technician always takes two measurements from different scales and records the average. In the Calon Scenario, measurement error could go up to 1 microgram on a single reading, and bias could go up to 8. The corresponding values were 4 and 5 for Kensington, and 8 and 1 for Alpine. These values were selected to provide maximal inferential power with three scenarios. Each scenario used different names for the two technicians. Participants responded on a 9-point scale, with the endpoints labeled [technician X] much closer. Appendix B gives the text for Calon.

After responding to the three scenario-specific accuracy questions, participants answered several directed questions about beliefs. These questions required that participants endorse statements about what happens to measurement error and bias when readings are averaged from either the same or different scales. A person's collection of responses to the directed questions constitutes an explicit belief pattern, which can then be used to refine and validate beliefs inferred from accuracy judgments. Two directed questions were used to establish belief patterns. The first tested beliefs about what happens to measurement error and bias when only one scale is used. The text is as follows:

> Suppose you have a scale like those used by the labs in the previous problems. The scale is subject to both measurement error and bias. As compared to weighing something just once, weighing an object multiple

times on this one scale, averaging the readings, and using the average as your estimate will ...

Participants completed this paragraph with one of several supplied statements about measurement error, and then with one of several similar statements about bias. These statements took the form: "tend to cause measurement errors (biases) to cancel out", "tend to have no effect on measurement errors (biases)", and "tend to cause measurement errors (biases) to add up". Participants also completed a similar paragraph depicting a scenario in which one reading is taken from each of multiple scales, with the final estimate again based on the average reading. This second paragraph supplied the same choice of statements about measurement error and bias.

The directed questions yield four specific beliefs about what happens to measurement error and bias when averages are based on either one scale or two. A person's responses can be represented by a sequence of four order relations that correspond to those in Table 1. For example, consider someone who states that using the same scale reduces measurement error and has no effect on bias, and that using different scales has no effect on measurement error and reduces bias. This combination of beliefs is represented by the notation (<, =, =, <), a variety of ETM. In addition to the directed questions, participants also explained their judgments in writing. Demographic data such as age, sex, and educational background were also collected.

Inferring Beliefs from Judgments

The term judgment sequence refers to a participant's accuracy judgments for the Calon, Kensington, and Alpine scenarios, in that order. Judgments from the 9-point response

scale were coded as favoring two measurements from the same scale (an 's' judgment), from different scales ('d'), or neither ('i', for indifferent). All analyses in this paper use these categorical judgments.[3] As there are three scenarios, there are 27 distinct judgment sequences possible. I use the term belief pattern to refer to a participant's beliefs about each of the four error relations depicted in Table 1. Since belief patterns are both inferred from judgment sequences, and elicited explicitly with the directed belief questions. There are three possible beliefs about each error relation (errors cancel, no effect, errors add up), which allows for 81 unique belief patterns.

For each judgment sequence, it is possible to list which of the 81 possible belief patterns might lead to that sequence. However, in a separate question (see Appendix C), 87% correctly indicated that averaging from the same scale has no effect on bias (i.e., Relation 2 holds as an equality). This constraint on Relation 2 reduces the number of possible belief patterns to 27. These patterns are listed in Table 2. Each pattern implies something about the judgments that a believer would make across the three scenarios. Pattern 1 is the normative model. Someone with this model would always judge two scales as more accurate because that way both bias and measurement error are reduced. Now consider Pattern 4, which describes one version of ETM. Use of the same scale reduces only measurement error while use of different scales reduces only bias. This model values redundancy when measurement error is substantially greater (implying an 's' judgment for Alpine), and nonredundancy when bias is substantially greater.

A potential difficulty in Table 2 is that a person might believe that both information collection methods reduce a given error type, but one does so better than the other. For

example, someone might believe in Pattern 1, but also believe that using the same scale reduces more measurement error than does using two scales. This set of beliefs could lead to an 's' judgment for Alpine, which has high potential for measurement error. An equally troublesome point is that a person might believe that one type of error is easier to reduce than the other. A Pattern 4 believer, for instance, might also believe that reductions in measurement error are typically greater than reductions in bias. This could lead to an 's' judgment for Kensington. To facilitate the analysis, I assume that people do not make such fine-grained distinctions. In other words, the symbols '<' and '>' imply the same proportionate amount of error reduction or increase wherever they appear. In most cases, relaxing this assumption will probably not change the implied judgments. The places where it could are marked by asterisks in Table 2. Similarly, occasionally no judgment emerges as the natural consequent of a belief pattern. These cases are marked by question marks.

There are several things to notice about the belief patterns and implied judgments in Table 2. First, many belief patterns reflect logical but unlikely possibilities. For example, it would be surprising to find believers in Pattern 19, which says that repeated measures from the same source increases overall error but measures from different sources reduces it. Interestingly, this "misguided" model produces normatively correct judgments in the three constructed scenarios. Second, any given judgment sequence (e.g., 'sss', 'ddd') is consistent with multiple belief patterns. Thus, the beliefs that underlie judgment can be only partially determined without the use of additional measures. Third, for certain judgment sequences it is possible to substantially reduce the list of potential underlying beliefs. For example, the sequence 'dds' matches only Patterns 4, 7, and 16 (and possibly 1). With some additional

measures and a bit of detective work, the judgment data could go a long way toward uncovering the beliefs that guide preferences for redundancy. Finally, only two belief Patterns, 4 and 7, fit the general ETM model from Table 1. Both imply a 'd' judgment for Calon and 's' for Alpine. If Patterns 1 and 16 can be ruled out, then 'd_s' responses can be attributed to some version of ETM.

Results

Reported Understanding.

On the 9-point ease of understanding scale, the terms measurement error, bias, and total deviation rated 8.20 (s.d. = 1.37), 8.15 (s.d. = 1.30), and 8.52 (s.d. = 1.00). In fact, each term was rated extremely easy to understand (rating of 9) by a majority of participants. For the vast majority, measurement error and bias are natural and intuitive concepts.

Judgment Sequences.

Out of 165 individual accuracy judgments (three for each of 55 participants), there were 80 (48.5%) 'd' judgments, 14 (8.5%) 'i' judgments, and 71 (43.0%) 's' judgments. The proportions choosing 'd', 'i', and 's' were as follows: Calon ('d' = .60, 'i' = .11, 's' = .29), Kensington (.49, .13, .38), and Alpine (.36, .02, .62). The proportion choosing 'd' varied significantly across the three scenarios (Cochran's $\chi^2(2) = 8.19$, $p < .05$; see Langley, 1970), indicating that when bias was relatively high, participants were more likely to respond that using different scales would be more accurate.

Table 3 divides judgment sequences into four nonoverlapping categories, including Nonredundant ('ddd' responses only), ETM-Consistent ('d_s'), Primarily Redundant (at least two 's' responses, no 'd's), and Other (everything else). Only 22% fell into the catch-all

Other category. The most common individual sequences were 'ddd', 'dds', and 'sss', which together account for 56% of responses. Table 3 also lists the belief patterns that are consistent with each observed judgment sequence. The asterisks indicate patterns that would be consistent assuming complex beliefs about the magnitudes and proportions of error reduction.

The individual sequences can be used to provide a more sensitive test of the manipulation. If the manipulation had no effect, all sequences with two 'd's and an 's', for instance, should occur equally often. In fact, the frequencies were 13 ('dds'), 1 ('dsd'), and 0 ('sdd') ($p < .05 \times 10^{-3}$, multinomial test). A comparison of 'd_s' ($n = 20$) and 's_d' ($n = 5$) gives a similar result ($p < .005$, binomial test). As ETM implies 'd_s', these results provide evidence for ETM as an intuitive model that guides choice. A similar comparison of 'dss' ($n = 5$), 'ssd' ($n = 5$), and 'sds' ($n = 1$) was nonsignificant ($p < .20$, multinomial test).

Explicit Belief Measures.

Table 4 shows overall response frequencies for each of the four directed belief items. Substantial majorities do appropriately believe that repeated use of the same scale averages out measurement error, and that the use of different scales averages out bias. This strong match between the normative model and reported beliefs verifies that most participants understood the question format as intended. A majority (58%) also correctly indicated that using the same scale multiple times will have no effect on the bias of the average. Finally, only 46% indicated that bias can be averaged out by using different scales.

Fifty-two of the 55 participants responded to all four directed belief questions, producing 24 unique patterns. These patterns were grouped into six non-overlapping

categories, listed in the first column of Table 5. The normative model and ETM were described earlier. The Apples & Oranges model derives its label from the common saying that two things are as different as apples and oranges, implying that they cannot be compared or, in this case, combined. This model holds that measurements from different scales cannot be combined productively, but that measurements from the same scale will tend to reduce error. The Apples & Apples model is the exact inverse of Apples & Oranges. It holds that there is no benefit to averaging multiple measurements from the same scale, but that averaging from different scales will reduce error. Finally, the No Difference model states that averaging has the same effects whether one scale is used or two.

Correspondence between Beliefs and Judgments.

In the previous two sections judgment sequences were categorized into four categories and explicit belief patterns into six. Were beliefs and judgments consistent? Table 5 crosstabulates belief patterns and judgment sequences. Consistent responses are indicated by asterisks. People with normative beliefs should always prefer nonredundant measurements ('ddd'), ETM implies 'd_s', Apples & Oranges implies Primarily Redundant responses, and Apples & Apples implies nonredundant responses. The No Difference and Other belief categories have no single best match. People described by No Difference may agree with the ordinal relationships, but still feel that using one scale or two affects the degree to which a given error type is reduced. Thus, it would be overly restrictive to assume that they must be indifferent in order to be consistent.

The analysis is restricted to the 36 participants in the top four rows. The base rates of the belief patterns and judgment sequences were used to calculate the probability that a

participant would be categorized in any one of the 24 cells, assuming independence.[4] The probability that a given participant would be categorized into one of the four matching cells is .27. Overall, 23 matches were observed ($p < .06 \times 10^{-4}$, binomial test). Each row in Table 5 was also tested individually, comparing the number of matches observed within that row with the number expected by chance. The p-values derived from the binomial tests are reported in the final column. Conditioning on measured beliefs, participants made judgments consistent with those beliefs. Notice also that the majority of participants with ETM-Consistent judgment sequences did in fact have ETM beliefs. This helps rule out Patterns 1 and 16 (see Table 2) as intuitive theories that underlie 'd_s' responses.

The results show that ETM beliefs underlie information source preference for many people. However, two specific varieties of ETM, Patterns 4 and 7, are still plausible. Both of these predict a 'd_s' judgment sequence. In support of Pattern 4, seven of the eleven participants classified as ETM for both beliefs and judgments reported Pattern 4 beliefs exactly, compared to only two for Pattern 7. Moreover, 'dds' ($n = 7$) responses seem more natural for Pattern 4, and 'dss' for Pattern 7 ($n = 2$). These results suggest that Pattern 4 beliefs underlie many instances of ETM preferences.

Demographic data.

Data were collected on age, sex, major area of study, and number of previous course in statistics. Table 6 is a correlation matrix including the number of 'd' judgments given across the three scenarios (D), the number of normative responses to the four directed belief items (B), sex, age, the number of courses taken in statistics, and major (science or nonscience). D and B are not as highly correlated as one might initially suppose ($r = .23$, $p =$

.11). This is not surprising given that incorrect beliefs can lead to a preference for nonredundancy, and nearly correct beliefs can lead a preference for redundancy. Table 7 displays several regression models with D as the dependent variable. The best model is one that includes Age and Major, but not the number of courses in statistics. A similar analysis was performed for B, and only Major was significant.

Overall, science majors chose the technician using two scales an average of 2.22 times, compared to 1.30 for nonscience majors ($t(53) = 2.47$, $p < .02$). In the explicit belief items, science majors reported normative principles an average of 3.50 times across the 4 items, compared to 2.34 for nonscience majors ($t(50) = 2.96$, $p < .005$). The success of science students is further underscored by the fact that they made up four of the five participants whose judgments and beliefs were both classified as normative.

Discussion

The results of Experiment 2 can be summarized as follows. First, consistent with Experiment 1 and past studies there was no general preference for redundant or nonredundant information sources. As expected, preference depends on the degree to which sources are prone to nonsystematic and systematic errors, and on one's intuitive theory about what happens to errors when estimates are aggregated. Second, judgments about the individual scenarios were highly consistent with explicit questions about beliefs. This suggests that people use intuitive theories of information to help decide which information sources to consult. Third, ETM was the most common belief pattern identified. Many participants reported ETM beliefs when asked directly, and their judgments in the three scenarios were

consistent with these beliefs. Some support was also found for the Normative model and for Apples & Oranges.

A fourth finding, that only science training is correlated with both the tendency to make normative judgments and report normative beliefs, deserves some discussion. Age predicted the number of normative judgments but not beliefs, and perhaps surprisingly the number of statistics courses taken was nonpredictive. The result for age suggests that people might learn over time that nonredundancy is valuable, but not learn why. There are several possible explanations for the success of science majors. First, some aspect of science training, perhaps laboratory work and experimentation, might foster adoption of the normative model. Second, those students more capable of understanding redundancy may be more likely to enter the sciences in the first place. Finally, participants with a science background might have benefited from the use of scientific stimuli. A different collection of stimuli might have favored a different group of participants. Additional work is needed to distinguish among these explanations.

As discussed earlier, Experiment 2 also asked participants to explain the reasoning behind their judgments in the three scenarios. The explanations were largely consistent with the judgments and reported beliefs, so a detailed analysis was omitted. As expected, most participants explained their choices in terms of bias and measurement error. Others highlighted additional factors that affect information seeking, such as the need for consistency in experimentation. As one participant explained,

> Because the scales are of the same type, which one you use does not matter. However, in any experiment it is sound procedure to keep procedures constant, in order to minimize the chance of extraneous random factors from corrupting the data.

Using multiple sources introduces new random factors in an experiment, which in the eyes of some tends to reduce accuracy. This type of argument can be interpreted in several ways. First, the proponent may be expressing a belief that using measurements from different sources increases one of the two error types. Such an opinion could reflect a failure to appreciate the statistical result that summing random variables increases variance while averaging reduces it. Second, the respondent might simply be misapplying the fundamental principle that one should minimize random factors impinging on an experiment (e.g., by keeping equipment clean, double-checking measurements, etc.). More correctly, one should increase the number of independent valid factors (e.g., instruments that produce uncorrelated errors), even if these new factors introduce new sources of error. Finally, some participants may have misunderstood the criterion. In comparing readings over time, one would in fact want to keep bias constant. For example, a dieter should use the same scale each day, because he or she is interested in the change in weight rather than a specific weight. The written explanations are not fine-grained enough to distinguish between these various reasons for liking consistency.

General Discussion

People often consult redundant sources of information, a strategy that in many situations is normatively suboptimal. Under certain conditions, people would send reconnaissance missions back to the same location, use the same medical test multiple times, and repeatedly weigh an object on the same scale. In each case, a nonredundant collection strategy (e.g., using different scales) dominates due to the expected lower correlation in forecast errors. The present paper traces the preference for redundancy to people's intuitive theories of information. The most popular theory is ETM, which holds that aggregating across sources cannot reduce nonsystematic errors. This belief leads people to incorrectly perceive a tradeoff between reducing nonsystematic errors with redundant sources and systematic errors with nonredundant sources. As a consequence, preferences for redundancy can be manipulated by varying the perceived expected sizes of the two kinds of error.

A potential concern with the present work is that the observed intuitive theories were invented "online" and do not reflect stable, enduring beliefs that apply to behavior outside the laboratory. Consistency both within and between individuals belies this interpretation. Experiment 2 measured beliefs in multiple ways and found a level of within-individual convergence that suggests that the observed intuitive theories are more than temporary inventions. This view is reinforced by the fact that participants converged on just several of the many possible intuitive theories. Nevertheless, one might expect that different situations would evoke different intuitive theories. A scientist, for example, might apply the normative model when using diagnostic instruments and ETM when collecting opinions from experts. One possibility is that people gravitate toward the normative model as they gain experience in

a domain. This raises the question of whether there is an initial default strategy that changes with life experience or domain expertise. For example, do people start out believing in Apples & Oranges and gradually adopt more sophisticated intuitive theories? The data hint at both general and domain-specific effects, since both age and science training predict preference for the nonredundant strategy when using scientific instruments to collect estimates.

A reasonable interpretation of the present findings is that people can learn to reason correctly about redundancy. This is consistent with research that shows that people have a repertoire of statistical and nonstatistical heuristics, and evoke one or the other depending on the features of the problem at hand (Nisbett, Krantz, Jepson, & Kunda, 1983). An example of a statistical heuristic is the law of large numbers (large samples are more reliable than small ones); whereas judging by representativeness (Kahneman & Tversky, 1972) is an example of a nonstatistical heuristic. People are more likely to employ statistical heuristics to the extent that the sampling process is salient, the event of interest is susceptible to random effects, and the culture encourages statistical reasoning (Nisbett et al., 1983). Use of statistical heuristics is also more likely as people acquire domain-specific expertise or general statistical training (Fong, Krantz, & Nisbett, 1986), which apparently enhances one's ability construct sample spaces and envision a distribution of possible events (Nisbett et al., 1983). The situation becomes more complicated when it comes to reasoning about redundancy and correlated error. Here, switching to a statistical mode of thought is not enough to generate normative behavior, because many people hold an incorrect statistical theory. Participant's written explanations clearly reveal reasoning that is simultaneously statistical and incorrect.

Appropriate statistical reasoning requires more than switching from a nonstatistical to a statistical mode of thought; it requires also that one's intuitive theory match the normatively correct one. For some principles, such as the law of large numbers, the main obstacle is simply recognizing when the statistical principle applies (Nisbett et al., 1983). For other principles, such as the rule that aggregating across sources reduces both systematic and nonsystematic errors, people's beliefs about how statistical information combines need to be reshaped. The current data suggest that it might help to encourage learning through active experimentation, as in some of the sciences, rather than through passive observation (Becker, 1996; Garfield, 1995; Klayman, 1988). Whether such training would foster an abstract intuitive model that could be applied across domains remains a topic for future study.

To the extent possible, this paper examined situations in which people collect and aggregate information in pursuit of accurate judgment. Most, if not all, normative models of judgment implicitly assume that accuracy is the sole objective. The present experiments were designed to eliminate, or at least severely attenuate, motivations other than accuracy. Even under these conditions, participants preferred to use redundant sources about half the time, although this proportion changed depending on the manipulation. There is, therefore, a cognitive account for why people like redundancy: there exists a fundamental mismatch between intuitive and normative theories of information.

It is reasonable to assume that accuracy is often only one of several motivations in a judgment task, and that other motivations may exacerbate the tendency to prefer redundancy. For example, consider a university administrator who champions a plan to add a business school to the campus. In forecasting the success of such a venture, the administrator may

want consensus, and therefore consult others who have in the past agreed with the administrator on a broad range of issues. However, this general past agreement is likely to have been caused in part by shared information, training, values, and perspective. In other words, a desire for agreement might bias information search in favor of redundant sources (Festinger, 1954; Frey, 1986), and hence various social motivations may amplify the cognitive factors that lead to a preference for redundancy.

Given that judgmental accuracy is limited by the quality of information upon which judgments are based, it is important to understand how, and how well, people collect information from multiple sources. The present research reveals systematic discrepancies between intuitive and normative theories of information. As a consequence, information search is inefficient, and judgment is less accurate than it otherwise might be. Even so, many people have sophisticated and nearly-correct intuitive theories. For example, ETM differs from the normative model only on the subtle point of what happens to nonsystematic error when one uses multiple sources of information. The data suggest that with the right training people might modify their intuitive theories, and as a consequence improve the quality of both information and judgment.

References

Ashton, A. H. (1986). Combining the judgments of experts: How many and which ones? Organizational Behavior and Human Decision Processes, 38, 405-414.

Becker, B. J. (1996). A look at the literature (and other resources) on teaching statistics. Journal of Educational and Behavioral Statistics, 21, 71-90.

Birnbaum, M. H., & Stegner, S. E. (1979). Source credibility in social judgment: Bias, expertise, and the judge's point of view. Journal of Personality and Social Psychology, 37, 48-74.

Clemen, R. T. (1989). Combining forecasts: A review and annotated bibliography. International Journal of Forecasting, 5, 559-609.

Clemen, R. T., & Winkler, R. L. (1985). Limits for the precision and value of information from dependent sources. Operations Research, 33, 427-442.

Clemen, R. T., & Winkler, R. L. (1986). Combining economic forecasts. Journal of Business and Economic Statistics, 4, 39-46.

Einhorn, H. J., Hogarth, R. M., & Klempner, E. (1977). Quality of group judgment. Psychological Bulletin, 84, 158-172.

Erev, I., Wallsten, T. S., & Budescu, D. V. (1994). Simultaneous over- and underconfidence: The role of error in judgment processes. Psychological Review, 101, 519-527.

Ferrell, W. R. (1985). Combining individual judgments. In G. Wright (Ed.), Behavioral Decision Making (pp. 111-145). New York: Plenum Press.

Festinger, L. (1954). A theory of social comparison processes. Human Relations, 7, 117-140.

Fong, G. T., Krantz, D. H., & Nisbett, R. E. (1986). The effects of statistical training on thinking about everyday problems. Cognitive Psychology, 18, 253-292.

Frey, D. (1986). Recent research on selective exposure to information. Advances in Experimental Social Psychology, 19, 41-80.

Garfield, J. (1995). How students learn statistics. International Statistical Review, 63, 25-34.

Goethals, G. R., & Nelson, R. E. (1973). Similarity in the influence process: The belief-value distinction. Journal of Personality and Social Psychology, 25, 117-122.

Goldberg, L. R. (1965). Diagnosticians versus diagnostic signs: The diagnosis of psychosis vs. neurosis from MMPI. Psychological Monographs, 79.

Gonzalez, R. (1994). When words speak louder than actions: Another's evaluations can appear more diagnostic than their decisions. Organizational Behavior and Human Decision Processes, 58, 214-245.

Hastie, R. (1986). Experimental evidence on group accuracy. In B. Grofman & G. Owen (Eds.), Decision Research (Vol. 2, pp. 129-157). Greenwich, CT: JAI Press.

Heath, C., & Gonzalez, R. (1995). Interaction with others increases decision confidence but not decision quality: Evidence against information collection views of interactive decision making. Organizational Behavior and Human Decision Processes, 61, 305-326.

Hogarth, R. M. (1978). A note on aggregating opinions. Organizational Behavior and Human Decision Processes, 21, 40-46.

Hogarth, R. M. (1989). On combining diagnostic 'forecasts': Thoughts and some evidence. International Journal of Forecasting, 5, 593-597.

Kahneman, D., & Tversky, A. (1972). Subjective probability: A judgment of representativeness. Cognitive Psychology, 3, 430-454.

Kahneman, D., & Tversky, A. (1973). On the psychology of prediction. Psychological Review, 80, 237-251.

Klayman, J. (1988). Cue discovery in probabilistic environments: Uncertainty and experimentation. Learning, Memory, and Cognition, 14(2), 317-330.

Larréché, J.-C., & Moinpour, R. (1983). Managerial judgment in marketing: The concept of expertise. Journal of Marketing Research, 20, 110-121.

Libby, R., & Blashfield, R. K. (1978). Performance of a composite as a function of the number of judges. Organizational Behavior and Human Decision Processes, 21, 121-129.

Maines, L. A. (1996). An experimental examination of subjective forecast combination. International Journal of Forecasting, 12, 223-234.

Makridakis, S., & Winkler, R. L. (1983). Averages of forecasts: Some empirical results. Management Science, 29, 987-996.

Mettee, D. R., & Smith, G. (1977). Social comparison and interpersonal attraction: The case for dissimilarity. In J. M. Suls & R. L. Miller (Eds.), Social Comparison Processes: Theoretical and Empirical Perspectives . Washington, DC: Hemisphere.

Nisbett, R. E., Krantz, D. H., Jepson, C., & Kunda, Z. (1983). The use of statistical heuristics in everyday reasoning. Psychological Review, 90, 339-363.

Overholser, J. C. (1994). The personality disorders: A review and critique of contemporary assessment strategies. Journal of Contemporary Psychotherapy, 24, 223-243.

Russ, R. C., Gold, J. A., & Stone, W. F. (1979). Attraction to a dissimilar stranger as a function of level of effectance arousal. Journal of Experimental Social Psychology, 15, 481-491.

Russ, R. C., Gold, J. A., & Stone, W. F. (1980). Opportunity for thought as a mediator of attraction to a dissimilar stranger: A further test of an information seeking interpretation. Journal of Experimental Social Psychology, 16, 562-572.

Sanders, F. (1963). On subjective probability forecasting. Journal of Applied Meteorology, 2(2), 191-201.

Slovic, P. (1966). Cue consistency and cue utilization in judgment. American Journal of Psychology, 79, 427-434.

Sniezek, J. A., & Buckley, T. (1995). Cueing and cognitive conflict in judge-advisor decision making. Organizational Behavior and Human Decision Processes, 62, 159-174.

Sniezek, J. A., & Henry, R. A. (1990). Revision, weighting, and commitment in consensus group judgment. Organizational Behavior and Human Decision Processes, 45, 66-84.

Soll, J. B. (1997). Lay theories of information: Beliefs about the value of redundancy. , University of Chicago, Chicago.

Staël Von Holstein, C.-A. S. (1971). An experiment in probabilistic weather forecasting. Journal of Applied Meteorology, 10, 635-645.

Stasson, M. F., & Hawkes, W. G. (1995). Effect of group performance on subsequent individual performance: Does influence generalize beyond the issues discussed by the group? Psychological Science, 305-307.

Wallsten, T. S., Budescu, D. V., & Erev, I. (1997). Evaluating and combining subjective probability estimates. Journal of Behavioral Decision Making, 10, 243-268.

Winkler, R. L., & Clemen, R. T. (1992). Sensitivity of weights in combining forecasts. Operations Research, 40, 609-614.

Zajonc, R. B. (1962). A note on group judgments and group size. Human Relations, 15, 177-180.

Appendix A

This appendix shows that

$$E(X_C\text{-}T)^2 = \frac{N_1 + N_2}{4} + \frac{S_1 + S_2}{4} + \frac{1}{2}\sqrt{S_1 S_2}\rho_S . \tag{A1}$$

Substituting $(X_1 + X_2)/2$ for X_C and manipulating terms yields

$$\frac{1}{4}E\left(X_1 - T\right)^2 + \frac{1}{4}E\left(X_2 - T\right)^2 + \frac{1}{2}E\left[\left(X_1 - T\right)\left(X_2 - T\right)\right] \tag{A2}$$

The first two expectations are the variances of the total error terms e_1 and e_2 from Eqs. 1 and 2. As nonsystematic and systematic errors are definitionally independent, the total error variance for a given source can be expressed as the sum of the component variances. Thus we have $E(X_1 - T)^2 = N_1 + S_1$ and $E(X_2 - T)^2 = N_2 + S_2$. Substituting these results into A2, rearranging terms, and expanding the third expectation in A2 yields

$$E(X_C - T)^2 =$$

$$\frac{N_1 + N_2}{4} + \frac{S_1 + S_2}{4} + \frac{E\left[\left(\left(X_1 - V_j\right) + \left(V_j - T\right)\right)\left(\left(X_1 - V_j\right) + \left(V_j - T\right)\right)\right]}{2} \tag{A3}$$

$$= \frac{N_1 + N_2}{4} + \frac{S_1 + S_2}{4} + \frac{E\left[\left(X_1 - V_j\right)\left(X_2 - V_k\right)\right]}{2}$$

$$+ \frac{E\left[\left(X_1 - V_j\right)\left(V_k - T\right)\right]}{2} + \frac{E\left[\left(V_j - T\right)\left(X_2 - V_k\right)\right]}{2} + \frac{E\left[\left(V_j - T\right)\left(V_k - T\right)\right]}{2} \tag{A4}$$

The first three expectations in A4 involve either two nonsystematic errors or one systematic and one nonsystematic. As each nonsystematic error is independent of all other errors, the expectations can be carried through as follows:

$$E(X_C - T)^2 = \frac{N_1 + N_2}{4} + \frac{S_1 + S_2}{4} + \frac{E(X_1 - V_j)E(X_2 - V_k)}{2}$$

$$+ \frac{E(X_1 - V_j)E(V_k - T)}{2} + \frac{E(V_j - T)E(X_2 - V_k)}{2} + \frac{E[(V_j - T)(V_k - T)]}{2} \qquad \text{(A5)}$$

$$= \frac{N_1 + N_2}{4} + \frac{S_1 + S_2}{4} + \frac{1}{2}\sqrt{S_1 S_2}\rho_S \qquad \text{(A6)}$$

To get from A5 to A6, note that by definition the individual errors all equal zero and thus drop out. Only $E((V_j - T)(V_k - T))$ remains. This term is simply the covariance of the two systematic errors, and can be recast in correlational terms using the relationship $Cov(X,Y) = \sigma_X \sigma_Y \rho_{XY}$.

Appendix B

The Calon Lab has many scales with which to weigh very small objects. All the scales are of the same type, and there is no way to tell whether one is more accurate than another. Scientists at Calon have determined that, for their type of scale, measurement error may be as much as 1 microgram and bias may be as much as 8 micrograms on a single reading from a given scale. Of course, figures for individual scales may vary.

As standard procedure, the Calon lab requires its technicians to weigh each object twice, and to record the average of the two measurements as the official estimate. For many years, Ashe and Birch have complied with this rule in different ways. Ashe chooses one scale randomly, puts the object on this scale twice in a row, and averages the two readings. Birch chooses two scales randomly, puts the object on each scale once, and averages the two readings. Over the long run, whose official estimates do you think come closer to the actual weights of the objects that they weigh? Recall that for a single reading from a given scale, measurement error may be as much as 1 microgram and bias may be as much as 8 micrograms.

Appendix C

Suppose you have a scale that tends to overestimate the weights of all objects by about 10 pounds. Sometimes it overestimates a little more, sometimes a little less, but on average it overestimates by 10 pounds. Suppose you put an object on this scale twice and take the average of the two readings. This average is more likely to

a) overestimate weight by more than 10 pounds

b) overestimate weight by less than 10 pounds

c) neither of the above is more likely

Footnotes

[1] Interestingly, when correlations between errors are very high, substantial error can be averaged out by assigning negative weight to the less accurate source (Clemen & Winkler, 1985). The relationship between the error variance of an appropriately weighted average and correlation is an inverted-u; error variance increases with correlation up to a point, after which it decreases rapidly. It can be shown that the threshold correlation at which this function peaks equals the ratio of the standard deviations of the error terms associated with the two unaveraged estimates. Soll (1997) showed that when errors are additively decomposable into a shared identical component and an uncorrelated unique component, the correlation in errors will never exceed this threshold. Empirical evidence occasionally yields correlations higher than the threshold, but the resulting weights are very unstable, generally leading to poorer performance than weighted averages that constrain the weights to be non-negative or simple rules such as equal weights (Winkler, personal communication; Winkler & Clemen, 1992). As a result, the promised lower error variance with very high correlations is not realized in practice.

[2] In the present treatment, nonsystematic errors are defined as independent. There may be cases that require a different treatment. For example, a person's random judgmental errors may be autocorrelated. This phenomenon could be modeled by decomposing random error even further into correlated and uncorrelated components, or by allowing for positive correlation in successive random errors. Whatever the modeling technique, the general normative prescription that one should prefer nonredundant information sources would still hold.

[3] The categorical judgments are preferred to the raw numerical responses because they more easily facilitate the assignment of intuitive labels to the judgment sequences. Analyses were performed on the raw numerical responses as well; the results are redundant with what is reported here. In the interest of parsimony, these analyses are omitted.

[4] Row marginals were normalized to ensure that probabilities summed to 1.

Table 1

Two Intuitive Theories

	Rela-tion #	Normative Model	ETM
Identical Information Sources	1	$N_C < N$	$N_C < N$
	2	$S_C = S$	S_C (= or >) S
Different Information Sources	3	$N_C < N$	N_C (= or >) N
	4	$S_C < S$	$S_C < S$

Table 2
Belief Patterns and Implied Judgment Sequences

Pattern #	Same Scale		Different Scales		Implied Judgments		
	Rel. 1 $N_C \gtrless N$	Rel. 2 $S_C \gtrless S$	Rel. 3 $N_C \gtrless N$	Rel. 4 $S_C \gtrless S$	Calon 1, 8	Kens'ton 4, 5	Alpine 8, 1
1	<	=	<	<	d	d*	d*
2	<	=	<	=	i*	i*	i*
3	<	=	<	>	s	s*	s*
4	<	=	=	<	d	d*	s
5	<	=	=	=	s	s	s
6	<	=	=	>	s	s	s
7	<	=	>	<	d	?	s
8	<	=	>	=	s	s	s
9	<	=	>	>	s	s	s
10	=	=	<	<	d	d	d
11	=	=	<	=	d	d	d
12	=	=	<	>	s	?	d
13	=	=	=	<	d	d	d
14	=	=	=	=	i	i	i
15	=	=	=	>	s	s	s
16	=	=	>	<	d	?	s
17	=	=	>	=	s	s	s
18	=	=	>	>	s	s	s
19	>	=	<	<	d	d	d
20	>	=	<	=	d	d	d
21	>	=	<	>	s	?	d
22	>	=	=	<	d	d	d
23	>	=	=	=	d	d	d
24	>	=	=	>	s	i*	d
25	>	=	>	<	d	d	d
26	>	=	>	=	i*	i*	i*
27	>	=	>	>	s	s*	s*

Note. Values beneath lab names indicate potential measurement error and bias, respectively. Asterisks indicate that implied judgment could differ given beliefs about magnitude of error reduction. Question marks indicate that there is no implied judgment.

Table 3

Frequency of Judgment Sequences and Consistent Beliefs

Label	Judgment Sequence	Frequency	Consistent Belief Patterns
Nonredundant	ddd	11	1, 2*, 10, 11, 13, 19, 20, 22, 23, 25, 26*
ETM-Consistent	dds	13	1*, 4, 7, 16
	dis	2	1*, 4, 7, 16
	dss	5	1*, 4, 7, 16
Primarily Redundant	sss	7	2*, 3, 5, 6, 8, 9, 15, 17, 18, 26*, 27
	sis	3	3*
	iss	2	--
Other	iii	1	2, 14, 26
	ssd	5	3*, 12, 21, 24*, 27*
	sds	1	--
	dsd	1	--
	did	1	--
	isd	1	--
	idd	1	--
	ids	1	--

Note. Asterisks indicate belief patterns which might be consistent given complex beliefs about the magnitudes and proportions of error reduction.

Table 4

Response Frequencies from Directed Belief Questions

	Cancels Out	No Effect	Adds Up
Use Same Scale Multiple Times			
Effect on Measurement error	43	9	3
	(.78)	(.16)	(.05)
Effect on Bias	11	30	11
	(.21)	(.58)	(.21)
Use Different Scales			
Effect on Measurement error	24	13	15
	(.46)	(.25)	(.29)
Effect on Bias	39	7	9
	(.71)	(.13)	(.16)

Note. Values in parentheses indicate proportions, within each row, reporting each response. Three subjects answered only two of the four items.

Table 5
Correspondence Between Beliefs and Judgments

Beliefs	Judgment Sequence: Always Nonredundant 'ddd'	ETM-Consistent 'd_s'	Primarily Redundant ≥ 2s, 0 d's	Other	Row Totals	p-value**
Normative	5*	3	1	--	**9** (.17)	.017
ETM	--	11*	1	3	**15** (.29)	.004
Apples & Oranges	--	1	4*	2	**7** (.13)	.054
Apples & Apples	3*	--	--	2	**5** (.10)	.052
No Difference	--	1	5	1	**7** (.13)	--
Other	2	3	1	3	**9** (.17)	--
Column Totals	**10** (.19)	**19** (.37)	**12** (.23)	**11** (.21)	**52**	

Note. * Indicates matched beliefs and judgments. Numbers in parentheses indicate proportion of all 52 subjects within each column and row. ** P-values derived from binomial test for first four belief patterns, and an exact test for the Other category.

Table 6

Pearson Correlation Matrix with Demographic Variables

	1	2	3	4	5	6
1. D	--					
2. B	0.227	--				
3. SEX	0.019	0.109	--			
4. AGE	0.345**	0.072	0.155	--		
5. STATS	0.221	0.039	0.123	0.562***	--	
6. MAJOR	0.321*	0.386**	0.123	0.120	0.040	--

Note. SEX = 1 if male, 0 if female; STATS = Number of courses in statistics; MAJOR = 1 if science major, 0 otherwise. * $p < .05$, ** $p < .01$, *** $p < .001$, all tests two-tailed.

Table 7

Regression of Number of Nonredundant Judgments on Various Predictors

	coefficient	standard coefficient	p	R^2
1. CONSTANT	-0.994	0.000	0.287	
2. AGE	0.116	0.345	0.010	0.119
1. CONSTANT	-0.857	0.000	0.430	
2. AGE	0.109	0.323	0.045	
3. STATS	0.053	0.040	0.800	0.120
1. CONSTANT	-0.886	0.000	0.325	
2. AGE	0.105	0.311	0.016	
3. MAJOR	0.811	0.283	0.028	0.198

Note. Major = 0 if nonscience student, 1 if science student.

Acknowledgment

This work is based on a doctoral thesis completed at the University of Chicago, under the guidance of Richard Larrick. Funding was provided by National Science Foundation Grant SBR-9409627, Decision, Risk, & Management Science Program. Financial support was also provided by an INSEAD research grant. Special thanks go to Richard Larrick and Joshua Klayman, for sage advice and input on previous drafts. Thanks also go to Chip Heath, Robin Hogarth, and my colleagues at the Center for Decision Research at the University of Chicago. I am indebted to Renuka Soll for assisting with the collection of data.